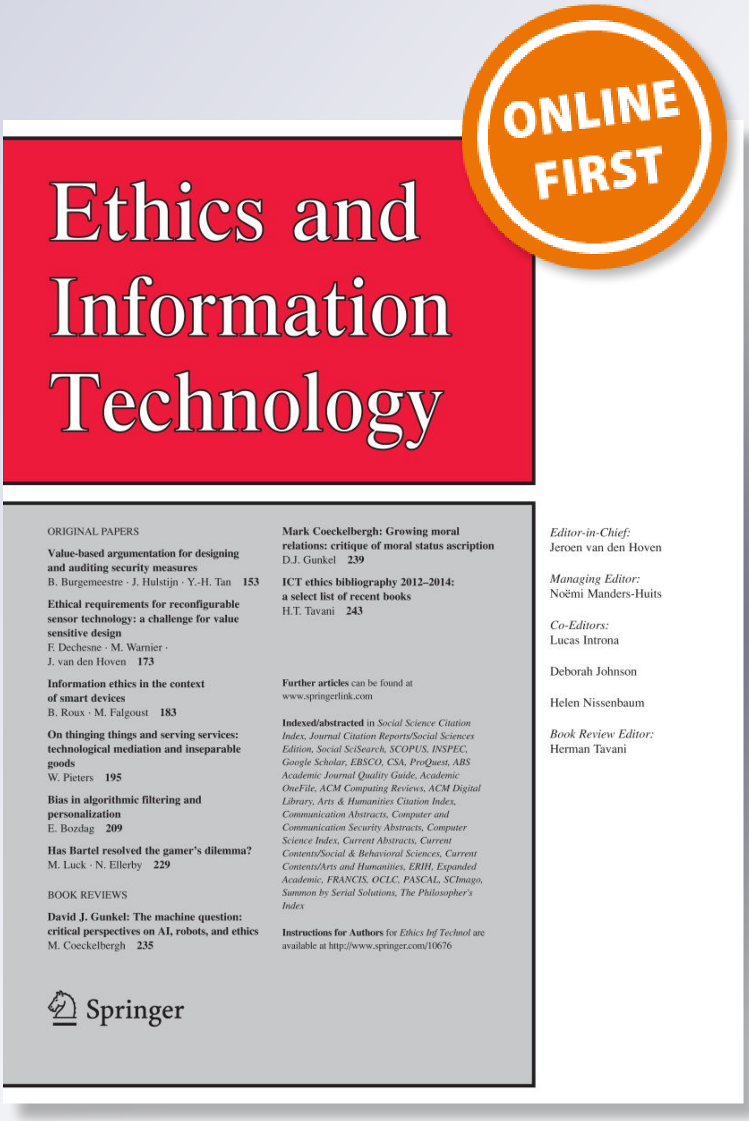
Mind the gap: responsible robotics and the problem of responsibility

David J. Gunkel

Ethics and Information Technology

ISSN 1388-1957

Ethics Inf Technol
DOI 10.1007/s10676-017-9428-2



ONLINE FIRST

Ethics and Information Technology

ORIGINAL PAPERS

Value-based argumentation for designing and auditing security measures
B. Burgemeestre · J. Hulstijn · Y.-H. Tan **153**

Ethical requirements for reconfigurable sensor technology: a challenge for value sensitive design
F. Dechesne · M. Warnier · J. van den Hoven **173**

Information ethics in the context of smart devices
B. Roux · M. Falgoust **183**

On thinging things and serving services: technological mediation and inseparable goods
W. Pieters **195**

Bias in algorithmic filtering and personalization
E. Bozdag **209**

Has Bartel resolved the gamer's dilemma?
M. Luck · N. Ellerby **229**

BOOK REVIEWS

David J. Gunkel: The machine question: critical perspectives on AI, robots, and ethics
M. Coeckelbergh **235**

Mark Coeckelbergh: Growing moral relations: critique of moral status ascription
D.J. Gunkel **239**

ICT ethics bibliography 2012–2014: a select list of recent books
H.T. Tavani **243**

Further articles can be found at www.springerlink.com

Indexed/abstracted in *Social Science Citation Index, Journal Citation Reports/Social Sciences Edition, Social SciSearch, SCOPUS, INSPEC, Google Scholar, EBSCO, CSA, ProQuest, ABS Academic Journal Quality Guide, Academic OneFile, ACM Computing Reviews, ACM Digital Library, Arts & Humanities Citation Index, Communication Abstracts, Computer and Communication Security Abstracts, Computer Science Index, Current Abstracts, Current Contents/Social & Behavioral Sciences, Current Contents/Arts and Humanities, ERIH, Expanded Academic, FRANCIS, OCLC, PASCAL, SCImago, Summon by Serial Solutions, The Philosopher's Index*

Instructions for Authors for *Ethics Inf Technol* are available at http://www.springer.com/10676

Springer

Springer

Ethics Inf Technol
DOI 10.1007/s10676-017-9428-2

ORIGINAL PAPER

Mind the gap: responsible robotics and the problem of responsibility

David J. Gunkel[1]

Abstract The task of this essay is to respond to the question concerning robots and responsibility—to answer for the way that we understand, debate, and decide who or what is able to answer for decisions and actions undertaken by increasingly interactive, autonomous, and sociable mechanisms. The analysis proceeds through three steps or movements. (1) It begins by critically examining the instrumental theory of technology, which determines the way one typically deals with and responds to the question of responsibility when it involves technology. (2) It then considers three instances where recent innovations in robotics challenge this standard operating procedure by opening gaps in the usual way of assigning responsibility. The innovations considered in this section include: autonomous technology, machine learning, and social robots. (3) The essay concludes by evaluating the three different responses—instrumentalism 2.0, machine ethics, and hybrid responsibility—that have been made in face of these difficulties in an effort to map out the opportunities and challenges of and for responsible robotics.

Keywords Robot · Robotics · Ethics · Machine ethics · Technology · Responsibility · Philosophy

Introduction

However it comes to be defined and characterized, "responsible robotics" is about responsibility of and for emerging technology. But "the concept of responsibility," as Riceour (2007, p.11) pointed out in his eponymously titled essay, is anything but clear and well-defined. Although the classical juridical usage of the term, which dates back to the nineteenth century, seems rather well-established—with "responsibility" characterized in terms of both civil and penal obligations (either the obligation to compensate for harms or the obligation to submit to punishment)—the philosophical concept is confused and somewhat vague.

> In the first place, we are surprised that a term with such a firm sense on the juridical plane should be of such recent origin and not really well established within the philosophical tradition. Next, the current proliferation and dispersion of uses of this term is puzzling, especially because they go well beyond the limits established for its juridical use. The adjective 'responsible' can complement a wide variety of things: you are responsible for the consequences of your acts, but also responsible for others' actions to the extent that they were done under your charge or care…In these diffuse uses the reference to obligation has not disappeared, it has become the obligation to fulfill certain duties, to assume certain burdens, to carry out certain commitments (Riceour 2007, pp. 11–12).

Riceour (2007, p. 12) traces this sense of the word through its etymology (hence the subtitle to the essay "A Semantic Analysis") to "the polysemia of the verb 'to respond'," which denotes "to answer for…." or "to respond to… (a question, an appeal, an injunction, etc.)." It is in this sense of the word that the question concerning responsibility has come to be associated with robotics. One of the principal issues for responsible robotics, if not the principal issue, is to decide who or what can be or should be

✉ David J. Gunkel
dgunkel@niu.edu
http://gunkelweb.com

[1] Northern Illinois University, Dekalb, IL 60115, USA

Published online: 19 July 2017

responsible for the consequences of decisions and actions instituted by robots or robotic systems? Who or what, in other words, can or should assume the obligations—the burden or duty—of answering for what a robot does or does not do?

The task of this essay is *to respond to* the question concerning robots and responsibility—*to answer for* the way that we understand, debate, and decide who or what is able *to answer for* decisions and actions undertaken by increasingly autonomous, interactive, and sociable mechanisms. In order to get at this, the following will proceed through three steps or movements. (1) I begin by critically examining the instrumental theory of technology, which determines the way one typically deals with and responds to the question of responsibility when it involves technology. (2) I then consider three instances where recent innovations challenge this standard operating procedure by opening gaps in the usual way of assigning responsibility. The innovations considered in this section include: autonomous technology, machine learning, and social robots. (3) I conclude by evaluating the three different responses—instrumentalism 2.0, machine ethics, and hybrid responsibility—that have been made in face of these difficulties in an effort to map out the opportunities and challenges of and for responsible robotics. The analysis is designed to be critical and not normative. The goal of the effort, therefore, is not to condemn instrumentalism per se but (1) to diagnose the challenges the instrumentalist way of thinking is now under due to recent innovations in information technology and (2) to evaluate the range of possible responses that can be made in the face of these challenges.[1]

Default settings

When it comes to the question of responsibility regarding technology, the matter seems rather clear and indisputable. "Morality," as Hall (2001, p. 2) points out, "rests on human shoulders, and if machines changed the ease with which things were done, they did not change responsibility for doing them. People have always been the only 'moral agents.'" This seemingly intuitive and common sense response is persuasive precisely because it is structured and informed by the answer that is typically provided for the question concerning technology. "We ask the question concerning technology," Heidegger (1977, pp. 4–5) writes, "when we ask what it is. Everyone knows the two statements that answer our question. One says: Technology is a means to an end. The other says: Technology is a human activity. The two definitions of technology belong together. For to posit ends and procure and utilize the means to them is a human activity." According to Heidegger's analysis, the presumed role and function of any kind of technology—whether it be a simple hand tool, jet airliner, or a sophisticated robot—is that it is a means employed by human users for specific ends. Heidegger calls this particular characterization of technology "the instrumental definition" and indicates that it forms what is considered to be the "correct" understanding of any kind of technological contrivance.

As Feenberg (1991, p. 5) summarizes it, "The instrumentalist theory offers the most widely accepted view of technology. It is based on the common sense idea that technologies are 'tools' standing ready to serve the purposes of users." And because a tool or instrument "is deemed 'neutral,' without valuative content of its own" a technological artifact is evaluated not in and of itself, but on the basis of the particular employments that have been decided by its human designer or user. Consequently, technology is only a means to an end; it is not and does not have an end in its own right. "Technical devices," as Lyotard (1993, p. 33) writes, "originated as prosthetic aids for the human organs or as physiological systems whose function it is to receive data or condition the context. They follow a principle, and it is the principle of optimal performance: maximizing output (the information or modification obtained) and minimizing input (the energy expended in the process). Technology is therefore a game pertaining not to the true, the just, or the beautiful, etc., but to efficiency: a technical 'move' is 'good' when it does better and/or expends less energy than another." According to Lyotard's analysis, a technological device, whether it be a cork screw, a clock, or a digital computer, is a mere instrument of human action. It therefore does not in and of itself participate in the big questions of truth, justice, or beauty. It is simply and indisputably about efficiency. A particular technological innovation is considered "good," if, and only if, it proves to be a more effective instrument (or means) to accomplishing a humanly defined end.

This formulation not only sounds level-headed and reasonable, it is one of the standard operating presumptions of

[1] This effort is informed by and consistent with the overall purpose and aim of philosophy, strictly speaking. Philosophers as different (and, at times, even antagonistic, especially to each other) as Heidegger (1962), Dennett (1996), Moore (2005), and Žižek (2006), have all, at one time or another, described philosophy as a critical endeavor that is more interested in developing questions than in providing definitive answers. "There are," as Žižek (2006, p. 137) describes it, "not only true or false solutions, there are also false questions. The task of philosophy is not to provide answers or solutions, but to submit to critical analysis the questions themselves, to make us see how the very way we perceive a problem is an obstacle to its solution." This is the task and objective of the essay—to identify the range of questions regarding responsibility that can and should be asked in the face of recent technological innovation. If, in the end, readers emerge from the experience with more questions—"more" not only in quantity but also (and more importantly) in terms of the quality of inquiry—then it will have been successful and achieved its end.

computer ethics. Although different definitions of "computer ethics" have circulated since Walter Maner first introduced the term in 1976, they all share a human-centered perspective that assigns responsibility to human designers and users. According to Deborah Johnson, who is credited with writing the field's agenda setting textbook, "computer ethics turns out to be the study of human beings and society—our goals and values, our norms of behavior, the way we organize ourselves and assign rights and responsibilities, and so on" (Johnson 1985, p. 6). Computers, she recognizes, often "instrumentalize" these human values and behaviors in innovative and challenging ways, but the bottom-line is and remains the way human beings design and use (or misuse) such technology. And Johnson has stuck to this conclusion even in the face of what appears to be increasingly sophisticated technological developments. "Computer systems," she writes in a more recent article, "are produced, distributed, and used by people engaged in social practices and meaningful pursuits. This is as true of current computer systems as it will be of future computer systems. No matter how independently, automatic, and interactive computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision" (Johnson 2006, 197). Understood in this way, computer systems, no matter how automatic, independent, or seemingly intelligent they may become, "are not and can never be (autonomous, independent) moral agents" (Johnson 2006, p. 203). They will, like all other technological artifacts, always be instruments of human value, decision making, and action.

According to the instrumentalist definition, therefore, any action undertaken via a technological system is ultimately the responsibility of some human agent—the designer of the system, the manufacturer of the equipment, or the end-user of the product. If something goes wrong with or someone is harmed by the mechanism, "some human is," as Goertzel (2002, p. 1) accurately describes it "to blame for setting the program up to do such a thing." Consequently, holding a robot responsible for the decisions it makes or the actions that it is instrumental in deploying is to make at least two errors. First, it is logically incorrect to ascribe agency to something that is and remains a mere object under our control. As Sullins (2006, p. 26) concludes by way of the investigations undertaken by Bringsjord (2007), computers and robots "will never do anything they are not programmed to perform" and as a result "are incapable of becoming moral agents now or in the future." This insight is a variant of one of the objections noted by Alan Turing in his agenda-setting paper on machine intelligence: "Our most detailed information of Babbage's Analytical Engine," Turing (1999, p. 50) wrote, "comes from a memoir by Lady Lovelace. In it she states, 'The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform' (her italics)." This objection—what Turing called "Lady Lovelace's Objection"—has often been deployed as the basis for denying independent agency or autonomy to computers, robots, and other mechanisms. Such instruments, it is argued, only do what we have programmed them to perform. Since we are the ones who deliberately design, develop, and deploy these mechanisms—or as Bryson (2010, p. 65) describes it, "there would be no robots on this planet if it weren't for deliberate human decisions to create them"—there is always a human that is if not *in the loop* then at least *on the loop*.

Second there are moral problems. That is, holding a robotic mechanism or system culpable would not only be illogical but also irresponsible. This is because ascribing moral responsibility to machines, as Siponen (2004, p. 286) argues, would allow one to "start blaming computers for our mistakes. In other words, we can claim that 'I didn't do it—it was a computer error', while ignoring the fact that the software has been programmed by people to 'behave in certain ways', and thus people may have caused this error either incidentally or intentionally (or users have otherwise contributed to the cause of this error)." This line of thinking has been codified in the popular adage, "It's a poor carpenter who blames his tools." In other words, when something goes wrong or a mistake is made in situations involving the application of technology, it is the operator of the tool and not the tool itself that should be blamed. "By endowing technology with the attributes of autonomous agency," Mowshowitz (2008, p. 271) argues, "human beings are ethically sidelined. Individuals are relieved of responsibility. The suggestion of being in the grip of irresistible forces provides an excuse of rejecting responsibility for oneself and others." This maneuver, what Nissenbaum (1996, p. 35) terms "the computer as scapegoat," is understandable but problematic, insofar as it allows human designers, developers, or users to deflect or avoid taking responsibility for their actions by assigning accountability to what is a mere object. "Most of us," Nissenbaum (1996, p. 35) argues, "can recall a time when someone (perhaps ourselves) offered the excuse that it was the computer's fault—the bank clerk explaining an error, the ticket agent excusing lost bookings, the student justifying a late paper. Although the practice of blaming a computer, on the face of it, appears reasonable and even felicitous, it is a barrier to accountability because, having found one explanation for an error or injury, the further role and responsibility of human agents tend to be underestimated—even sometimes ignored. As a result, no one is called upon to answer for an error or injury." It is precisely for this reason that Johnson and Miller (2008, 124) argue that "it is dangerous to conceptualize computer systems as autonomous moral agents." Assigning responsibility to the technology not only sidelines human involvement and activity, but leaves questions of responsibility untethered from their assumed proper attachment to human decision making and action.

The robot apocalypse

The instrumental theory not only sounds reasonable, it is obviously useful. It is, one might say, instrumental for responding to the opportunities and challenges of increasingly complex technological systems and devices. This is because the theory has been successfully applied not only to simple devices like corkscrews, toothbrushes, and garden hoses but also sophisticated technologies, like computers, smart phones, drones, etc. But all of that appears to be increasingly questionable or problematic, precisely because of a number of recent innovations that effectively challenge the operational limits of the instrumentalist theory.

Machine != Tool

The instrumental theory is a rather blunt instrument, reducing all technology, irrespective of design, construction, or operation, to the ontological status of a tool or instrument. "Tool," however, does not necessarily encompass everything technological and does not, therefore, exhaust all possibilities. There are also *machines*. Although "experts in mechanics," as Marx (1977, p. 493) pointed out, often confuse these two concepts calling "tools simple machines and machines complex tools," there is an important and crucial difference between the two. Indication of this essential difference can be found in a brief parenthetical remark offered by Heidegger in "The Question Concerning Technology." "Here it would be appropriate," Heidegger (1977, p. 17) writes in reference to his use of the word "machine" to characterize a jet airliner, "to discuss Hegel's definition of the machine as autonomous tool [selbständigen Werkzeug]." What Heidegger references, without supplying the full citation, are Hegel's 1805-07 Jena Lectures, in which "machine" had been defined as a tool that is self-sufficient, self-reliant, and independent. Although Heidegger immediately dismisses this alternative as something that is not appropriate to his way of questioning technology, it is taken up and given sustained consideration by Langdon Winner in his book-length investigation of *Autonomous Technology*.

> To be autonomous is to be self-governing, independent, not ruled by an external law or force. In the metaphysics of Immanuel Kant, autonomy refers to the fundamental condition of free will—the capacity of the will to follow moral laws which it gives to itself. Kant opposes this idea to "heteronomy," the rule of the will by external laws, namely the deterministic laws of nature. In this light the very mention of autonomous technology raises an unsettling irony, for the expected relationship of subject and object is exactly reversed. We are now reading all of the propositions backwards. To say that technology is autonomous is to say that it is nonheteronomous, not governed by an external law. And what is the external law that is appropriate to technology? Human will, it would seem (Winner 1977, p. 16).

"Autonomous technology," therefore, refers to technological devices that directly contravene the instrumental definition by deliberately contesting and relocating the assignment of agency. Such mechanisms are not mere tools to be used by human beings but occupy, in one way or another, the place of human agent. As Marx (1977, p. 495) described it, "the machine, therefore, is a mechanism that, after being set in motion, performs with its tools the same operations as the worker formerly did with similar tools." Understood in this way, the machine does not occupy the place of the tool used by the worker; it takes the place of the worker him/herself. This is precisely why the question concerning automation, or robots in the work place, is not merely able to be explained away as the implementation of new and better tools but raises concerns over the replacement of human workers or what has been called, beginning with the work of Keynes (2010, p. 325), "technological unemployment."

Perhaps the best example of the difference Marx describes is available to us with the self-driving car or autonomous vehicle. The autonomous vehicle, whether the Google Car or one of its competitors, is not designed for and intended to replace the automobile. It is, in its design, function, and materials, the same kind of instrument that we currently utilize for the purpose of personal transportation. The autonomous vehicle, therefore, does not replace the instrument of transportation (the car); it is intended to replace (or at least significantly displace) the driver. This difference was recently acknowledged by the National Highway Traffic Safety Administration (NHTSA), which in a 4 February 2016 letter to Google, stated that the company's Self Driving System (SDS) could legitimately be considered the legal driver of the vehicle: "As a foundational starting point for the interpretations below, NHTSA will interpret 'driver' in the context of Google's described motor vehicle design as referring to the SDS, and not to any of the vehicle occupants" (Ross 2016). Although this decision is only an interpretation of existing law, the NHTSA explicitly states that it will "consider initiating rulemaking to address whether the definition of 'driver' in Section 571.3 [of the current US Federal statute, 49 U.S.C. Chapter 301] should be updated in response to changing circumstances" (Hemmersbaugh 2016). Similar proposals have been floated in efforts to deal with work-place automation. In a highly publicized draft document submitted to the European Parliament in May of 2016, for instance, it was argued that "sophisticated autonomous robots" ("machines" in Marx's terminology) be considered "electronic persons" with "specific rights and obligations" for the purposes of contending with the challenges of technological unemployment, tax policy, and legal liability. Although the proposed legislation did not pass as originally written, it represents recognition on the part of lawmakers that recent innovations in robotics challenge the way we typically respond to and answer for questions regarding responsibility.

The instrumentalist theory works by making the assumption that all technologies—irrespective of design, implementation, or sophistication—are a tool of human action. Hammer, computer, or UAV (unmanned aerial vehicle), they are all just instruments that are used more or less effectively and/or correctly by human beings. But this way of thinking does not cover everything; there are also *machines*. Machines, as Marx (following Hegel's initial suggestions) recognized, occupy another ontological position. They are not instruments to be used (more or less efficiently) by a human agent; they are designed and implemented to take the place of the human agent. Consequently, machines—like the self-driving automobile and other forms of what Winner calls "autonomous technology"—challenge the explanatory capability of the instrumentalist theory, presenting us with technologies that are intentionally designed and deployed to be something other. Pointing this out, however, does not mean that the instrumental theory is on this account refuted *tout court*. There are and will continue to be mechanisms understood and utilized as tools to be manipulated by human users (i.e., lawn mowers, cork screws, telephones, etc.). The point is that the instrumentalist perspective, no matter how useful and seemingly correct in some circumstances for answering for some technological devices, does not exhaust all possibilities for all kinds of devices. The theory has its limits.

Learning algorithms

The instrumental theory, for all its notable success handling different kinds of technology (for a critical examination of examples of these "success stories," see Heidegger 1977; Feenberg 1991; Nissenbaum 1996 and; Johnson 2006), appears to be unable to contend with recent developments in machine learning. Consider, for example, Google DeepMind's AlphaGo and Microsoft's Tay.ai.[2] Although not physically embodied mechanisms, both AlphaGo and Tay demonstrate the "responsibility gap" (Matthias 2004) that is opening up in the wake of recent innovations in machine learning. AlphaGo, as Google DeepMind (2016) explains, "combines Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play." In other words, AlphaGo does not play the game of Go by following a set of cleverly designed moves feed into it by human programmers. It is designed to formulate its own instructions. Although less is known about the inner workings of Tay, Microsoft explains that the system "has been built by mining relevant public data," i.e. training its neural networks on anonymized data obtained from social media, and was designed to evolve its behavior from interacting with users on social networks like Twitter, Kik, and GroupMe (Microsoft 2016). What both systems have in common is that the engineers who designed and built them have little or no idea what the systems will eventually do once they are in operation. As Thore Graepel, one of the creators of AlphaGo, has explained: "Although we have programmed this machine to play, we have no idea what moves it will come up with. Its moves are an emergent phenomenon from the training. We just create the data sets and the training algorithms. But the moves it then comes up with are out of our hands" (Metz 2016). Consequently, machine learning systems, like AlphaGo, are intentionally designed to do

[2] Because of the recent proliferation of and popularity surrounding connectionist architecture, neural networks, and machine learning, there are numerous examples from which one could select, including natural language generation (NLG) algorithms, black box trading, computational creativity, self-driving vehicles, and autonomous weapons. In fact, one might have expected this essay to have focused on the latter—autonomous weapons—mainly because of the way the responsibility gap, or what has also been called "the accountability gap," has been positioned, addressed, and documented in the literature on this subject (Arkin 2009; Asaro 2012; Beard 2014; Hammond 2015; Krishnan 2009; Lokhorst and van den Hoven 2012; Schulzke 2013; Sharkey 2012; Sparrow 2007; Sullins 2010). I have, however, made the deliberate decision to employ other, perhaps more mundane, examples like AlphaGo and Tay.ai. And I have done so for two reasons. First, questions concerning machine autonomy and responsibility, although important for and well-documented in the literature concerning autonomous weapons, is something that is not (and should not be) limited to weapon systems. Recognizing this fact requires that we explicitly identify and consider other domains where these question appear and are relevant—domains where the issues might be less dramatic but no less significant. Second, and more importantly, I wanted to deal with technologies that are actually in operation and not under development. Despite its popularity in investigations of machine agency and responsibility, autonomous weapons are still somewhat speculative and in development. Rather than address *what might happen* with technologies that could be developed and deployed, I wanted to address *what has happened* with technologies that are already here and in operation.

things that their programmers cannot anticipate, completely control, or answer for. In other words, we now have autonomous (or at least semi-autonomous) computer systems that in one way or another have "a mind of their own." And this is where things get interesting, especially when it comes to questions of responsibility.

AlphaGo was designed to play Go, and it proved its ability by beating an expert human player. So who won? Who gets the accolade? Who actually beat Lee Sedol? Following the dictates of the instrumental theory of technology, actions undertaken with the computer would be attributed to the human programmers who initially designed the system and are capable of answering for what it does or does not do. But this explanation does not necessarily hold for an application like AlphaGo, which was deliberately created to do things that exceed the knowledge and control of its human designers. In fact, in most of the reporting on this landmark event, it is not Google or the engineers at DeepMind who are credited with the victory. It is AlphaGo. In published rankings, for instance, it is AlphaGo that is named as the number two player in the world (Go Ratings 2016). Things get even more complicated with Tay, Microsoft's foul-mouthed teenage AI, when one asks the question: Who is responsible for Tay's bigoted comments on Twitter? According to the standard instrumentalist way of thinking, we could blame the programmers at Microsoft, who designed the application to be able to do these things. But the programmers obviously did not set out to create a racist algorithm. Tay developed this reprehensible behavior by learning from interactions with human users on the Internet. So how did Microsoft answer for this? How did they explain and assign responsibility?

Initially a company spokesperson—in damage-control mode—sent out an email to *Wired, The Washington Post*, and other news organizations, that sought to blame the victim. "The AI chatbot Tay," the spokesperson explained, "is a machine learning project, designed for human engagement. It is as much a social and cultural experiment, as it is technical. Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways. As a result, we have taken Tay offline and are making adjustments" (Risely 2016). According to Microsoft, it is not the programmers or the corporation who are responsible for the hate speech. It is the fault of the users (or some users) who interacted with Tay and taught her to be a bigot. Tay's racism, in other word, is our fault. Later, on Friday the 25th of March, Peter Lee, VP of Microsoft Research, posted the following apology on the Official Microsoft Blog: "As many of you know by now, on Wednesday we launched a chatbot called Tay. We are deeply sorry for the unintended offensive and hurtful tweets from Tay, which do not represent who we are or what we stand for, nor how we designed Tay. Tay is now offline and we'll look to bring Tay back only when we are confident we can better anticipate malicious intent that conflicts with our principles and values" (Lee 2016). But this apology is also frustratingly unsatisfying or interesting (it all depends on how you look at it). According to Lee's carefully worded explanation, Microsoft is only responsible for not *anticipating* the bad outcome; it does not take responsibility or answer for the offensive Tweets. For Lee, it is Tay who (or "that," and words matter here) is named and recognized as the source of the "wildly inappropriate and reprehensible words and images" (Lee 2016). And since Tay is a kind of "minor" (a teenage AI) under the protection of her parent corporation, Microsoft needed to step-in, apologize for their "daughter's" bad behavior, and put Tay in a time out.

Although the extent to which one might assign "agency" and "responsibility" to these mechanisms remains a contested issue, what is not debated is the fact that the rules of the game have changed and that there is a widening "responsibility gap."

> Presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, by the machine itself. This is what we call machine learning. Traditionally we hold either the operator/manufacture of the machine responsible for the consequences of its operation or "nobody" (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume responsibility for them (Matthias 2004, p. 177).

In other words, the instrumental definition of technology, which had effectively tethered machine action to human agency and responsibility, no longer adequately applies to mechanisms that have been deliberately designed to operate and exhibit some form, no matter how rudimentary, of independent action or autonomous decision making. Contrary to the instrumentalist way of thinking, we now have mechanisms that are designed to do things that exceed our control and our ability to respond or to answer for them. But let's be clear as to what this means. What has been demonstrated is not that a machine, like AlphGo or Tay, is or should be considered a moral agent and held solely accountable for the decisions it makes or the actions it deploys. That would be going too far, and it would be inattentive to the actual results that have been obtained. In fact, if we return to Riceour (2007) and follow his lead, which suggests that responsibility be understood as the "ability to respond," it is clear that both AlphaGo and Tay lack this capability. If we should, for instance, want to know more about the moves that AlphaGo made in its historic game against Lee Sedol, AlphaGo can certainly be asked about it. But the algorithm will have nothing to say in response. In fact, it was the responsibility of the human programmers and observers to respond on behalf of AlphaGo and to explain the significance and impact of its behavior. But what this does indicate is that machine learning systems like AlphaGo and Tay.ai introduce complications into the instrumentalist way of assigning and dealing with responsibility. They might not be moral agents in their own right (not yet at least), but their design and operation effectively challenge the standard instrumentalist theory and open up fissures in the way responsibility comes to be decided, assigned, and formulated.

Social robots

In July of 2014 the world got its first look at Jibo. Who or what is Jibo? That is an interesting and important question. In a promotional video that was designed to raise capital investment through pre-orders, social robotics pioneer Cynthia Breazeal introduced Jibo with the following explanation: "This is your car. This is your house. This is your toothbrush. These are your things. But these [and the camera zooms into a family photograph] are the things that matter. And somewhere in between is this guy. Introducing Jibo, the world's first family robot" (Jibo 2014). Whether explicitly recognized as such or not, this promotional video leverages a crucial ontological distinction that Derrida (2005, p. 80) calls the difference between "who" and "what." On the side of "what" we have those things that are mere objects—our car, our house, and our toothbrush. According to the instrumental theory of technology, these things are mere instruments that do not have any independent moral status whatsoever. We might worry about the impact that the car's emissions have on the environment (or perhaps stated more precisely, on the health and well-being of the other human beings who share this planet with us), but the car itself is not an object of moral concern. On the other side there are, as the video describes it "those things that matter." These things are not "things," strictly speaking, but are the other persons who count as socially and morally significant Others. They are those others to whom we are obligated and in the face of which we bear certain duties or responsibilities. Unlike the car, the house, or the toothbrush, these other persons have moral status and can be benefitted or harmed by our decisions and actions.

Jibo, we are told, occupies a place that is situated somewhere in between *what* are mere things and *who* really matters. Consequently Jibo is not just another instrument, like the automobile or toothbrush. But he/she/it (and the choice of pronoun is not unimportant) is also not quite another member of the family pictured in the photograph. Jibo inhabits a place in between these two ontological categories.[3] This is, it should be noted, not unprecedented. We are already familiar with other entities that occupy a similar ambivalent social position, like the family dog. In fact animals, which since the time of Descartes have been the other of the machine (Gunkel 2007), provide a good precedent for understanding the changing nature of social responsibility in the face of social robots, like Jibo. "Looking at state of the art technology," Darling (2012, p. 1) writes, "our robots are nowhere close to the intelligence and complexity of humans or animals, nor will they reach this stage in the near future. And yet, while it seems far-fetched for a robot's legal status to differ from that of a toaster, there is already a notable difference in how we interact with certain types of robotic objects." This occurs, Darling continues, principally due to our tendencies to anthropomorphize things by projecting into them cognitive capabilities, emotions, and motivations that do not necessarily exist in the mechanism per se. But it is this emotional reaction that necessitates new forms of obligation in the face of social robots. "Given that many people already feel strongly about state-of-the-art social robot 'abuse,' it may soon become more widely perceived as out of line with our social values to treat robotic companions in a way that we would not treat our pets" (Darling 2012, p. 1).

This insight is not just a theoretical possibility; it has been demonstrated in empirical investigations. The computer as social actor (CASA) studies undertaken by Reeves and Nass (1996), for example, demonstrated that human users will accord computers social standing similar to that of another human person and that this occurs as a product of the extrinsic social interaction, irrespective of the actual internal composition (or "being" as Heidegger would say) of the object in question. These results, which were obtained in numerous empirical studies with human subjects over several years, have been independently verified

[3] Just to be clear, the problem with social robots is not that they are or might be capable of becoming moral subjects. The problem is that they are neither instruments nor moral subjects. They occupy an in-between position that effectively blurs the boundary that had typically separated the one from the other. The problem, then, is not that social robots might achieve moral status equal to or on par with human beings. That remains a topic of and for science fiction. The problem is that social robots complicate the way one decides who has moral status and what does not, which is a more difficult/interesting philosophical question. For more on this subject, see Coeckelbergh (2012), Gunkel (2012), and Floridi (2013).

in two recent experiments with robots, one reported in the *International Journal of Social Robotics* (Rosenthal-von der Pütten et al. 2013), where researchers found that human subjects respond emotionally to robots and express empathic concern for machines irrespective of knowledge concerning the actual ontological status of the mechanism, and another that used physiological evidence, documented by electroencephalography, of the ability of humans to empathize with what appears to be simulated "robot pain" (Suzuki et al. 2015).

Jibo, and other social robots like it, are not science fiction. They are already or will soon be in our lives and in our homes. As Breazeal (2002, p. 1) describes it, "a sociable robot is able to communicate and interact with us, understand and even relate to us, in a personal way. It should be able to understand us and itself in social terms. We, in turn, should be able to understand it in the same social terms—to be able to relate to it and to empathize with it…In short, a sociable robot is socially intelligent in a human-like way, and interacting with it is like interacting with another person." In the face of these socially situated and interactive entities we are going to have to decide whether they are mere things like our car, our house, and our toothbrush; someone who matters and to whom we bear responsibility, like another member of the family; or something altogether different that is situated in between the one and the other, like our pet dog. In whatever way this comes to be decided, however, these artifacts will undoubtedly challenge our understanding of responsibility and the way we typically distinguish between *who* is to be considered another social subject and *what* is a mere instrument or object. Again, let's be absolutely clear about things. Social robots, like Jibo, are not (at least not at this particular moment in history) considered to be either a moral agent or a moral patient (on this distinction, see Gunkel 2012 and; Floridi 2013). But Jibo is also something more than a mere tool or object. It occupies what is arguably an ambivalent in-between position that complicates the usual way of thinking about and assigning moral status. Social robots like Jibo, then, are designed and situated in such a way that they do not fit the available ontological and axiological categories; their very existence already complicates the usual way of sorting things into "who" or "what."

Responding to responsibility gaps

What we have discovered, therefore, are situations where our theory of technology—a theory that has considerable history behind it and that has been determined to be as applicable to simple hand tools as it is to complex technological systems—encounters significant difficulties responding to or answering for recent developments with autonomous technology, machine learning, and social robots. In the face of these challenges, there are at least three possible responses.

Instrumentalism 2.0

We can try to respond as we typically have responded, treating these recent innovations in artificial intelligence and robotics as mere instruments or tools. Joanna Bryson makes a persuasive case for this approach in her provocatively titled essay "Robots Should be Slaves": "My thesis is that robots should be built, marketed and considered legally as slaves, not companion peers" (Bryson 2010, p. 63). Although this might sound harsh, the argument is persuasive, precisely because it draws on and is underwritten by the instrumental theory of technology. This decision has both advantages and disadvantages. On the positive side, it reaffirms human exceptionalism, making it absolutely clear that it is only the human being who possess rights and responsibilities. Technologies, no matter how sophisticated they become, are and will continue to be mere tools of human action, nothing more. "We design, manufacture, own and operate robots" Bryson (2010, p. 65) writes. "They are entirely our responsibility. We determine their goals and behaviour, either directly or indirectly through specifying their intelligence, or even more indirectly by specifying how they acquire their own intelligence." Furthermore, this line of reasoning seems to be entirely consistent with current legal structures and decisions. "As a tool for use by human beings," Gladden (2016, p. 184) concludes, "questions of legal responsibility for any harmful actions performed by such a robot revolve around well-established questions of product liability for design defects (Calverley 2008, p. 533; Datteri 2013) on the part of its producer, professional malpractice on the part of its human operator, and, at a more generalized level, political responsibility for those legislative and licensing bodies that allowed such devices to be created and used."

But this approach, for all its usefulness, has at least two problems. First, in strictly applying and enforcing the instrumental theory, we might inadvertently restrict technological innovation and the development of responsible governance. If, for example, we hold developers responsible for the unintended consequences of robots that have been designed with learning capabilities, this could lead engineers and manufactures to be rather conservative with the development and commercialization of new technology in an effort to protect themselves from culpability. Had the engineers at Microsoft been assigned responsibility for the hate speech produced by Tay, it is very possible that they and the corporation for whom they worked (or the legal department within the corporation) would think twice before releasing such technology into the wild. This might be, it could be argued, a positive development, similar to the safety measures and product testing requirements that are currently employed in the development of pharmaceuticals, transportation systems, and other industries. But it could also restrict and hinder robust development of machine learning systems and capabilities.

There is a similar situation with self-driving cars and the evolution of governance. As the NHTSA explicitly noted in its letter to Google, trying to assign the responsibility of "driver" to some human being riding in the autonomously driven vehicle would be both practically and legally inaccurate. "No human occupant of the SDV [Self Driven Vehicle] could meet the definition of "driver" in Section 571.3 given Google's described motor vehicle design, even if it were possible for a human occupant to determine the location of Google's steering control system, and sit 'immediately behind' it, that human occupant would not be capable of actually driving the vehicle as described by Google. If no human occupant of the vehicle can actually drive the vehicle, it is more reasonable to identify the 'driver' as whatever (as opposed to whoever) is doing the driving" (Hemmersbaugh 2016). For this reason, "accepting an AI as a legal driver eases the government's rule-writing process" (Ross 2016) by making existing law applicable to recent changes in automotive technology.

Second, strict application of the instrumental theory to robots, as Bryson directly acknowledges, produces a new class of instrumental servant or slave, what we might call, following Gunkel (2012, p. 86) "slavery 2.0." The problem here, as Brooks (2002) insightfully points out, is not with the concept of "slavery" per se (we should not, in other words, get hung up on words); the problem has to do with the kind of robotic mechanisms to which this term comes to be applied:

> Fortunately we are not doomed to create a race of slaves that is unethical to have as slaves. Our refrigerators work twenty-four hours a day seven days a week, and we do not feel the slightest moral concern for them. We will make many robots that are equally unemotional, unconscious, and unempathetic. We will use them as slaves just as we use our dishwashers, vacuum cleaners, and automobiles today. But those that we make more intelligent, that we give emotions to, and that we empathize with, will be a problem. We had better be careful just what we build, because we might end up liking them, and then we will be morally responsible for their well-being. Sort of like children (Brooks 2002, p. 195).

As Brooks explains, our refrigerators work tireless on our behalf and "we do not feel the slightest moral concern for them." But things will be very different with social robots, like Jibo, that invite and are intentionally designed for emotional investment and attachment.

Contra Brooks, however, it seems we are already at this point with things that are (at least metaphorically) as cold and impersonal as the refrigerator. As Singer (2009, p. 338) and Garreau (2007) have reported, US soldiers in Iraq and Afghanistan have formed surprisingly close personal bonds with their units' Packbots, giving them names, awarding them battlefield promotions, risking their own lives to protect that of the robot, and even mourning their death. This happens, Singer explains, as a product of the way the mechanism is situated within the unit and the role that it plays in battlefield operations. And it happens in direct opposition to what otherwise sounds like good common sense: They are just technologies—instruments or tools that work on our behalf and feel nothing. This "correction," in fact, is part and parcel of the problem. This is because the difficulties with strict reassertion of the instrumental theory has little or nothing to do with speculation about what the mechanism may or may not feel. The problem is with the kind of social environment it produces. As Kant (1963, p. 239) argued concerning indirect duties to non-human animals: Animal abuse is wrong, not because of how the animal might feel (which is, according to Kant's strict epistemological restrictions, forever and already inaccessible to us), but because of the adverse effect such action would have on other human beings and society as a whole. In other words, applying the instrumental theory to these new kinds of mechanism and affordances, although seemingly reasonable and useful, could have potentially devastating consequences for us, our world, and the other entities we encounter here.

Machine ethics

Conversely, we can entertain the possibility of what has been called "machine ethics" just as we had previously done for other non-human entities, like animals (Singer 1975). And there has, in fact, been a number of recent proposals addressing this opportunity. Wallach and Allen (2009, p. 4), for example, not only predict that "there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight" but use this fact as justification for developing "moral machines," advanced technological systems that are able to respond to morally challenging situations. Anderson and Anderson (2011) take things one step further. They not only identify a pressing need to consider the moral responsibilities and capabilities of increasingly autonomous systems but have even suggested that "computers might be better at following an ethical theory than most humans," because humans "tend to be inconsistent in their reasoning" and "have difficulty juggling the complexities of ethical decision-making" owing to the sheer volume of data that need to be taken into account and processed (Anderson and Anderson 2007, p. 5).

These proposals, it is important to point out, do not necessarily require that we first resolve the "big questions" of AGI (Artificial General Intelligence), robot sentience, or machine consciousness. As Wallach (2015, p. 242) points out, these kinds of machines need only be "functionally moral." That is, they can be designed to be "capable of making ethical determinations…even if they have little or no actual understanding of the tasks they perform." The precedent for this way of thinking can be found in corporate law and business ethics. Corporations are, according to both national and international law, legal persons (French 1979). They are considered "persons" (which is, we should recall, a moral classification and not an ontological category) not because they are conscious entities like we assume ourselves to be, but because social circumstances make it necessary to assign personhood to these artificial entities for the purposes of social organization and jurisprudence. Consequently, if entirely artificial and human fabricated entities, like Google or IBM, are legal persons with associated social responsibilities, it would be possible, it seems, to extend the same moral and legal considerations to an AI or robot like Google's DeepMind or IBM's Watson. The question, it is important to point out, is not whether these mechanisms are or could be "natural persons" with what is assumed to be "genuine" moral status; the question is whether it would make sense and be expedient, from both a legal and moral perspective, to treat these mechanisms as persons in the same way that we currently do for corporations, organizations and other human artifacts.

Once again, this decision sounds reasonable and justified. It extends both moral and legal responsibility to these other socially aware and interactive entities and recognizes, following the predictions of Wiener (1988, p. 16), that the social situation of the future will involve not just human-to-human interactions but relationships between humans and machines and machines and machines. But this shift in perspective also has significant costs. First, it requires that we rethink everything we thought we knew about ourselves, technology, and ethics. It entails that we learn to think beyond human exceptionalism, technological instrumentalism, and many of the other *-isms* that have helped us make sense of our world and our place in it. In effect, it calls for a thorough reconceptualization of who or what should be considered a legitimate center of moral concern and why.

Second, robots that are designed to follow rules and operate within the boundaries of some kind of programmed restraint, might turn out to be something other than what is typically recognized as a responsible agent. Winograd (1990, pp. 182–183), for example, warns against something he calls "the bureaucracy of mind," "where rules can be followed without interpretive judgments." "When a person," Winograd (1990, p. 183) argues, "views his or her job as the correct application of a set of rules (whether human-invoked or computer-based), there is a loss of personal responsibility or commitment. The 'I just follow the rules' of the bureaucratic clerk has its direct analog in 'That's what the knowledge base says.' The individual is not committed to appropriate results, but to faithful application of procedures." Coeckelbergh (2010, p. 236) paints a potentially more disturbing picture. For him, the problem is not the advent of "artificial bureaucrats" but "psychopathic robots." The term "psychopathy" has traditionally been used to name a kind of personality disorder characterized by an abnormal lack of empathy which is masked by an ability to appear normal in most social situations. The functional morality, like that specified by Anderson and Anderson and Wallach and Allen, intentionally designs and produces what are arguably "artificial psychopaths"—robots that have no capacity for empathy but which follow rules and in doing so can appear to behave in morally appropriate ways. These psychopathic machines would, Coeckelbergh (2010, p. 236) argues, "follow rules but act without fear, compassion, care, and love. This lack of emotion would render them non-moral agents—i.e. agents that follow rules without being moved by moral concerns—and they would even lack the capacity to discern what is of value. They would be morally blind."[4]

Efforts in "machine ethics" (or whatever other nomenclature comes to be utilized to name this development) effectively seek to widen the circle of moral subjects to include what had been previously excluded and marginalized as mere neutral instruments of human action. This is, it is important to note, not some blanket statement that would turn everything that was a tool into a moral subject. It is the recognition, following Marx, that not everything technological is reducible to a tool and that some devices—what Marx called "machines" and what Winner calls "autonomous technology"—might need to be

[4] There is some debate concerning this matter. What Coeckelbergh (2010, p. 236) calls "psychopathy"— e.g. "follow rules but act without fear, compassion, care, and love"—Arkin (2009) celebrates as a considerable improvement in moral processing and decision making. Here is how Sharkey (2012, p. 121) characterizes Arkin's efforts to develop an "artificial conscience" for robotic soldiers: "It turns out that the plan for this conscience is to create a mathematical decision space consisting of constraints, represented as prohibitions and obligations derived from the laws of war and rules of engagement (Arkin 2009). Essentially this consists of a bunch of complex conditionals (if-then statements)….Arkin believes that a robot could be more ethical than a human because its ethics are strictly programmed into it, and it has no emotional involvement with the action." For more on this debate and the effect it has on moral consideration, see Gunkel (2012).

programmed in such a way as to behave reasonably and responsibly for the sake of respecting human individuals and communities. This proposal has the obvious advantage of responding to moral intuitions: if it is the machine that is making the decision and taking action in the world with little or no direct human oversight, it would only make sense to hold it accountable (or at least partially accountable) for the actions it deploys and to design it with some form of constraint in order to control for possible bad outcomes. But doing so has considerable costs. Even if we bracket the questions of AGI, super intelligence, and machine consciousness; designing robotic systems that follow prescribed rules might provide the right kind of external behaviors but the motivations for doing so might be lacking. "Even if," Sharkey (2012, p. 121) writes in a consideration of autonomous weapons, "a robot was fully equipped with all the rules from the Laws of War, and had, by some mysterious means, a way of making the same discriminations as humans make, it could not be ethical in the same way as is an ethical human. Ask any judge what they think about blindly following rules and laws." Consequently, what we actually get from these efforts might be something very different from (and maybe even worse than) what we had hoped to achieve.

Hybrid responsibility

Finally, we can try to balance these two opposing positions by taking an intermediate hybrid approach, distributing responsibility across a network of interacting human and machine components. Hanson (2009, p. 91), for instance, introduces something he calls "extended agency theory," which is itself a kind of extension/elaboration of the "actor-network theory" initially developed by Latour (2005). According to Hanson, who takes what appears to be a practical and entirely pragmatic view of things, robot responsibility is still undecided and, for that reason, one should be careful not to go too far in speculating about things. "Possible future development of automated systems and new ways of thinking about responsibility will spawn plausible arguments for the moral responsibility of non-human agents. For the present, however, questions about the mental qualities of robots and computers make it unwise to go this far" (Hanson 2009, p. 94). Instead, Hanson suggests that this problem may be resolved by considering various theories of "joint responsibility," where "moral agency is distributed over both human and technological artifacts" (Hanson 2009, p. 94).

This proposal, which can be seen as a kind of elaboration of Nissenbaum's (1996) "many hands" thesis, has been gaining traction, especially because it appears to be able to deal with and respond to complexity. According to van de Poel et al. (2012, pp. 49–50): "When engineering structures fail or an engineering disaster occurs, the question who is to be held responsible is often asked. However, in complex engineering projects it is often quite difficult to pinpoint responsibility." As an example of this, the authors point to an investigation of 100 international shipping accidents undertaking by Wagenaar and Groenewegen (1987, p. 596): "Accidents appear to be the result of highly complex coincidences which could rarely be foreseen by the people involved. The unpredictability is due to the large number of causes and by the spread of the information over the participants." For van de Poel et al. (2012, pp. 50–51), however, a more informative example can be obtained from the problem of climate change. "We think climate change is a typical example of a many hands problem because it is a phenomenon that is very complex, in which a large number of individuals are causally involved, but in which the role of individuals in isolation is rather small. In such situations, it is usually very difficult to pinpoint individual responsibility. Climate change is also a good example of how technology might contribute to the occurrence of the problem of many hands because technology obviously plays a major role in climate change, both as cause and as a possible remedy".

Extended agency theory, therefore, moves away from the anthropocentric individualism of enlightenment thought, what Hanson (2009, p. 98) calls "moral individualism," and introduces an ethic that is more in-line with recent innovations in ecological thinking:

> When the subject is perceived more as a verb than a noun—a way of combining different entities in different ways to engage in various activities—the distinction between Self and Other loses both clarity and significance. When human individuals realize that they do not act alone but together with other people and things in extended agencies, they are more likely to appreciate the mutual dependency of all the participants for their common well-being. The notion of joint responsibility associated with this frame of mind is more conducive than moral individualism to constructive engagement with other people, with technology, and with the environment in general (Hanson 2009, p. 98).

Similar proposals has been advanced and advocated by Deborah Johnson and Peter Paul Verbeek for dealing with innovation in information technology. "When computer systems behave," Johnson (2006, p. 202) writes, "there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user." "I will," Verbeek (2011, p. 13) argues, "defend the thesis that ethics should be approached as a matter of human-technological associations. When taking the notion of technological mediation seriously, claiming that technologies are human agents would be as inadequate as claiming that ethics is a solely human affair." For both Johnson and Verbeek, responsibility is something distributed across a network of interacting components and these networks include not just other human persons, but organizations, natural objects, and technologies.

This hybrid formulation—what Verbeek calls "the ethics of things" and Hanson terms "extended agency theory"—has advantages and disadvantages. To its credit, this approach appears to be attentive to the exigencies of life in the twenty-first century. None of us, in fact, make decisions or act in a vacuum; we are always and already tangled up in networks of interactive elements that complicate the assignment of responsibility and decisions concerning who or what is able to answer for what comes to pass. And these networks have always included others—not only other human beings but institutions, organizations, and even technological components like the robots and algorithms that increasingly help organize and dispense with social activity. This combined approach, however, still requires that someone decide and answer for what aspects of responsibility belong to the machine and what should be retained for or attributed to the other elements in the network. In other words, "extended agency theory," will still need to decide *who* is able to answer for a decision or action and *what* can be considered a mere instrument (Derrida 2005, p. 80).

Furthermore, these decisions are (for better or worse) often flexible and variable, allowing one part of the network to protect itself from culpability by instrumentalizing its role and deflecting responsibility and the obligation to respond elsewhere. This occurred, for example, during the Nuremberg trials at the end of World War II, when low-level functionaries tried to deflect responsibility up the chain of command by claiming that they "were just following orders." But the deflection can also move in the opposite direction, as was the case with the prisoner abuse scandal at the Abu Ghraib prison in Iraq during the presidency of George W. Bush. In this situation, individuals in the upper echelon of the network deflected responsibility down the chain of command by arguing that the documented abuse was not ordered by the administration but was the autonomous action of a "few bad apples" in the enlisted ranks. Finally, there can be situations where no one or nothing is accountable for anything. In this case, moral and legal responsibility is disseminated across the elements of the network in such a way that no one person, institution, or technology is culpable or held responsible. This is precisely what happened in the wake of the 2008 financial crisis. The bundling and reselling of mortgage-backed securities was considered to be so complex and dispersed across the network that in the final analysis no one was able to be identified as being responsible for the collapse.

Conclusions

From the beginning our concern has been the concept and exigencies of responsibility. Usually efforts to decide the question of responsibility in the face of technology is not a problem, precisely because the instrumental theory assigns responsibility to the human being and defines technology as nothing more than a mere tool or instrument. It is, therefore, the human being who is responsible for responding or answering for what the machine does nor does not do (or perhaps more accurately stated, what comes to be done or not done through the instrumentality of the mechanism). This way of thinking has worked rather well, with little or no significant friction, for over 2500 years, and it holds considerable promise for application to the project of responsible robotics. But, as we have seen, recent innovations in technology—autonomous machines, learning algorithms, and social robots—challenge the instrumental theory by opening up what Matthias (2004) calls "responsibility gaps."

In response to these challenges—in an effort to close or at least remediate the gap—we have considered three alternatives. On the one side, there is strict application of the instrumental theory, which would restrict all questions of responsibility to human beings and define robots, no matter how sophisticated their design and operations, as nothing more than tools or instruments of human decision making and action. On the other side, there are efforts to assign some level of moral agency to machines. Even if robots are not (at least for now) able to be full moral subjects, they can, it is argued, be functionally responsible. Though such "responsibility" is only a kind of "quasi-responsibility" (Stahl 2006), this way of thinking assigns the ability to respond to the mechanism. And situated somewhere in between these two opposing positions, is a kind of intermediate option that distributes responsibility (and the ability to respond) across a network of interacting components, some human and some entirely otherwise.

These three options clearly define a spectrum of possible responses with each mode of response having its own particular advantages and disadvantages. Consequently, how we—individually but also as a collective—decide to respond to these opportunities and challenges will have a profound effect on the way we conceptualize our place in the world, who we decide to include in the community of moral subjects, and what we exclude from such consideration and why. But no matter how it is decided, it is a decision—quite literally a cut that institutes difference and makes a difference. We are, therefore, responsible both for deciding who or even what is a moral subject and, in the process, for determining the current state and future possibility of and for responsible robotics.

References

Anderson, M., & Anderson, S. L. (2007). The status of machine ethics: A report from the AAAI symposium. *Minds & Machines, 17*(1), 1–10.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge: Cambridge University Press.

Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press.

Asaro, P. (2012). On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross, 94*(886), 687–709.

Beard, J. M. (2014). Autonomous weapons and human responsibilities. *Georgetown Journal of International Law, 45*(1), 617–681.

Breazeal, C. L. (2004). *Designing sociable robots*. Cambridge, MA: MIT Press.

Bringsjord, S. (2007). Ethical robots: The future can heed us. *AI & Society, 22*(4), 539–550.

Brooks, R. A. (2002). *Flesh and machines: How robots will change us*. New York: Pantheon Books.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). Amsterdam: John Benjamins.

Calverley, D. J. (2008). Imaging a non-biological machine as a legal person. *AI & Society, 22*(4), 523–537.

Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. New York: Palgrave Macmillan.

Committee on Legal Affairs. Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics. European Parliament, 2016. http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN.

Darling, K. (2012). Extending legal protection to social robots. *IEEE Spectrum*. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots.

Datteri, E. (2013). Predicting the long-term effects of human-robot interaction: A reflection on responsibility in medical robotics. *Science and Engineering Ethics, 19*(1), 139–160.

Dennett, D. C. (1996). *Kinds of minds: Toward and understanding of consciousness*. New York: Perseus Books.

Derrida, J. (2005). *Paper machine* (trans. by R. Bowlby). Stanford, CA: Stanford University Press.

Feenberg, A. (1991). *Critical theory of technology*. New York: Oxford University Press.

Floridi, L. (2013). *The ethic of information*. Oxford: Oxford University Press.

French, P. (1979). The corporation as a moral person. *American Philosophical Quarterly, 16*(3), 207–215.

Garreau, J. (2007). Bots on the Ground: In the Field of Battle (or Even Above it), Robots are a Soldier's Best Friend. *The Washington Post*, Retrieved May 6, 2007, from http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html.

Gladden, M. E. (2016). The diffuse intelligent other: An ontology of nonlocalizable robots as moral and legal actors. In M. Nørskov (Ed.), *Social robots: Boundaries, potential, challenges* (pp. 177–198). Burlington, VT: Ashgate.

Go Ratings. (2016). https://www.goratings.org/.

Goertzel, B. (2002). Thoughts on AI morality. *Dynamical Psychology: An International, Interdisciplinary Journal of Complex Mental Processes*, May 2002. http://www.goertzel.org/dynapsyc/2002/AIMorality.htm.

Google DeepMind. (2016). AlphaGo. https://deepmind.com/alpha-go.html.

Gunkel, D. J. (2007). Thinking otherwise: Ethics, technology and other subjects. *Ethics and Information Technology, 9*(3), 165–177.

Gunkel, D. J. (2012). *The machine question: Critical perspectives on ai, robots and ethics*. Cambridge, MA: MIT Press.

Hall, J. S. (2001). Ethics for machines. *KurzweilAI.net*. http://www.kurzweilai.net/ethics-for-machines.

Hammond, D. N. (2015). Autonomous weapons and the problem of state accountability. *Chicago Journal of International Law, 15*(2), 652–687.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(1), 91–99.

Heidegger, M. (1962). *Being and time* (trans. by John Macquarrie and Edward Robinson). New York: Harper and Row.

Heidegger, M. (1977). *The Question concerning technology and other essays* (trans. by William Lovitt). New York: Harper and Row.

Hemmersbaugh, P. A. NHTSA Letter to Chris Urmson, Director, Self-Driving Car Project, Google, Inc. https://isearch.nhtsa.gov/files/Google - compiled response to 12 Nov 15 interp request - 4 Feb 16 final.htm.

Jibo. (2014). https://www.jibo.com.

Johnson, D. G. (1985). *Computer ethics*. Upper Saddle River, NJ: Prentice Hall.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2–3), 123–133.

Kant, I. (1963). *Duties to animals and spirits. lectures on ethics* (trans. by L. Infield) (pp. 239–241). New York: Harper and Row.

Keynes, J. M. (2010). Economic possibilities for our grandchildren. In *Essays in persuasion* (pp. 321–334). New York: Palgrave Macmillan.

Krishnan, A. (2009). *Killer robots: Legality and ethicality of autonomous weapons*. Burlington: Ashgate.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.

Lee, P. Learning from Tay's introduction. *Official Microsoft Blog*, 25 March 2016. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Lokhorst, G. J., & van den Hoven, J. (2012). Responsibility for military robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robots* (pp. 145–155). Cambridge, MA: MIT Press.

Lyotard, J. F. (1993). *The postmodern condition: A report on knowledge* (trans. by Geoff Bennington and Brian Massumi). Minneapolis, MN: University of Minnesota Press.

Marx, K. (1977). *Capital* (trans. by Ben Fowkes). New York: Vintage Books.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Metz, C. Google's AI Wins a Pivotal Second Game in Match with Go Grandmaster. *Wired*, March 2016. http://www.wired.

com/2016/03/googles-ai-wins-pivotal-game-two-match-go-grandmaster/.

Microsoft. (2016). Meet Tay—Microsoft AI. Chatbot with Zero Chill. https://www.tay.ai/.

Moore, G. E. (2005). *Principia ethica*. New York: Barnes & Noble Books.

Mowshowitz, A. (2008). Technology as excuse for questionable ethics. *AI & Society, 22*(3), 271–282.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics, 2*(1), 25–42.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.

Riceour, P. (2007). *Reflections on the just* (trans. by David Pellauer). Chicago: University of Chicago Press.

Risely, J. (2016). Microsoft's Millennial Chatbot Tay.ai Pulled Offline After Internet Teaches Her Racism. *GeekWire*. http://www.geekwire.com/2016/even-robot-teens-impressionable-microsofts-tay-ai-pulled-internet-teaches-racism/.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics, 5*(1), 17–34.

Ross, P. E. (2016). A google car can qualify as a legal driver. *IEEE Spectrum*. http://spectrum.ieee.org/cars-that-think/transportation/self-driving/an-ai-can-legally-be-defined-as-a-cars-driver.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*(2), 203–219.

Sharkey, N. (2012). Killing made easy: From joysticks to politics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robots* (pp. 111–128). Cambridge, MA: MIT Press.

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals*. New York: New York Review Book.

Singer, P. W. (2009). *Wired for war: The robotics revolution and conflict in the twenty-first century*. New York: Penguin Books.

Siponen, M. (2004). A pragmatic evaluation of the theory of information ethics. *Ethics and Information Technology, 6*(4), 279–290.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8*(4), 205–213.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*(12), 23–30.

Sullins, J. P. (2010). Robowarfare: Can robots be more ethical than humans on the battlefield? *Ethics and Information Technology, 12*(3), 263–275.

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, *5*(1), 15924. doi:10.1038/srep15924.

Turing, A. (1999). Computing machinery and intelligence. In P. A. Meyer (Ed.), *Computer media and communication: A reader* (pp. 37–58). Oxford: Oxford University Press.

van de Poel, I., Nihle´n Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science Engineering Ethics, 18*(1), 49–67.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Wagenaar, W. A., & Groenewegen, J. (1987). Accidents at sea: Multiple causes and impossible consequences. *International Journal of Man-Machine Studies, 27*, 587–598.

Wallach, W. (2015). *A dangerous master: How to keep technology from slipping beyond our control*. New York: Basic Books.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Wiener, N. (1988). *The human use of human beings: Cybernetics and society*. Boston: Ad Capo Press.

Winner, L. (1977). *Autonomous technology: Technics-out-of-control as a theme in political thought*. Cambridge, MA: MIT Press.

Winograd. T. (1990). Thinking machines: Can there be? Are we? In D. Partridge & Y. Wilks (Eds.), *The foundations of artificial intelligence: A sourcebook* (pp. 167–189). Cambridge: Cambridge University Press.

Žižek, S. (2006). Philosophy, the "Unknown Knowns," and the public use of reason. *Topoi, 25*(1–2), 137–142.